SUMMARY OF THE DATA SCIENCE METHODOLOGY AS APPLIED TO THE PROBLEM

Firstly, data was imported to a dataframe using pandas library. Of the initial 37 variables, 22 of these which had a high percentage of null/unique values or which didn't hold any meaningful/useful information were dropped. The remaining data was cleaned and prepared for analysis by treating missing values/outliers and creating dummy variables for categorical variables. 'Select' values in the dataframe probably represents cases where the customer missed to select a value from a drop-down list and is treated as missing values. Further, the target variable was separated, and Test-train split of the data was done. The features were standardized using the standard scaler method in Scikitlearn.

Then, a logistic regression model was built in Python using the function GLM() under statsmodel library. This model which contained all the variables, underwent further variable-pruning via a hybrid approach – coarse tuning using RFE followed by manual fine-tuning based on the VIFs and p-values. After a rigorous and iterative manual feature selection process, the final model ended up with 13 variables and accuracy of 92%. On ensuring that there was no significant multi-collinearity in the model, the specificity and sensitivity was computed from the confusion matrix. The specificity and sensitivity were pretty good (~96% and 85% respectively) on the training dataset. To further authenticate the model, ROC curve was plotted on the train data. The area under the curve (AUC) came around 95% which was pretty good value. After plotting the accuracy, specificity and sensitivity for various cut-off probabilities, the optimal cut-off was around 0.2 and this value was chosen to be the threshold and got decent values of all the five metrics – Accuracy (~91%), Sensitivity (~85%), Specificity (~94%), Precision(~93%) and Recall(~85%). Lead score was also assigned in addition to the conversion prediction.

The model was deployed on the test data set after scaling the features. Predictions were made and metrics were re-run. Now, the scores were - Accuracy (~90%), Sensitivity (~84%) and Specificity (~94%). The area under the ROC curve was roughly 95% again which was indicative of a good model. Since both specificity and sensitivity values ranged high, the model would easily cater to the season-specific business requirement to identify all possible conversion leads or just reliable/accurate conversion leads.

Lead score was calculated on the entire model. Lead conversion rate for X Education as deduced by the model was 87%. To further exploit business opportunities, the features were sorted based on their importance (i.e. correlation with lead conversion). The top 5 features which had the highest probabilities of favouring lead conversion came out to be Tags_Lost to EINS, Tags_Closed by Horizzon, Tags_Will revert after reading the email, Tags_Busy and Lead Source_Welingak Website while the negatively correlated ones were Do Not Email, Tags_Ringing, Tags_switched off , Lead Quality_Not Sure and Lead Quality_Worst. To further enhance the conversion rates, all that the company needed to do was to nurture the leads with the former features and work on reducing the latter.